Module 19 Multivariate Analysis for Genetic data
Session 09 Discriminant Analysis

Jan Graffelman

jan.graffelman@upc.edu

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

28th Summer Institute in Statistical Genetics (SISG 2023)

Contents

Discriminant analysis: Aims

- Group separation
- Dimension reduction: from p variables to k discriminators with $k < p$.
- Classification of new cases

Discriminant Analysis: the data matrix

Ind.	X_1	X_2	$\cdots$	X_p	Group
1	X_{11}	X_{12}	$\cdots$	X_{1p}	1
2	X_{21}	X_{22}	$\cdots$	X_{2p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	$X_{n_1 1}$	$X_{n_1 2}$	$\cdots$	$X_{n_1 p}$	1
1	X_{11}	X_{12}	$\cdots$	X_{1p}	2
2	X_{21}	X_{22}	$\cdots$	X_{2p}	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_2	$X_{n_2 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_2 p}$	2
1	X_{11}	X_{12}	$\cdots$	X_{1p}	m
2	X_{21}	X_{22}	$\cdots$	X_{2p}	m
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_m	$X_{n_m 1}$	$X_{n_m 2}$	$\cdots$	$X_{n_m p}$	m

Some examples

- Given various biochemical measurements, is this person healthy or diseased?
- Given the variables of this wheat kernel, to which of the known varieties does it belong?
- ...
- One can distinguish between two-group and multiple group problems.

Two-group linear discriminant analysis

Criteria for designing a classification rule:

- small probability of misclassification
- take prevalence into account (prior probabilities)
- take the cost of misclassification into account

Two-group linear discriminant analysis

Some basic definitions:

- π_1 and π_2 represent population 1 and 2.
- $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ represent the multivariate probability densities for each population.
- $\Omega = R_1 \cup R_2$ is the partitioned sample space for outcome $\mathbf{x}$.
- If $\mathbf{x}$ falls in R_1, the case is classified as π_1, else in π_2.
- p_1 is the prior probability of pertaining to π_1, p_2 the prior probability of pertaining to π_2 (prevalence)
- Misclassification probabilities:
  1. $P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x}$
  2. $P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$

Cost matrix

		Predicted class	
		π_1	π_2
True	π_1	0	$c(2\|1)$
Class	π_2	$c(1\|2)$	0

$c(1|2)$ and $c(2|1)$ are not necessarily equal

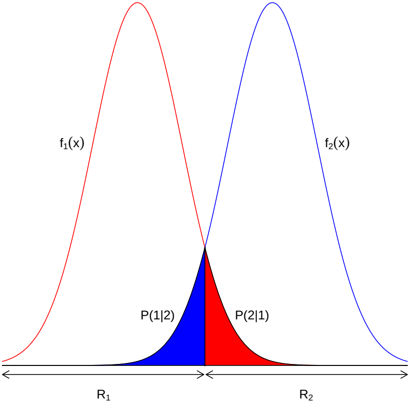
ECM = Expected Cost of Misclassification

$$P\left(\text{from } \pi_1 \cap \text{classified } \pi_2\right) = P\left(2|1\right) \cdot p_1$$

$$P\left(\text{from } \pi_2 \cap \text{classified } \pi_1\right) = P\left(1|2\right) \cdot p_2$$

$$\text{ECM} = c(1|2)P(1|2)p_2 + c(2|1)P(2|1)p_1$$

Classification rule: minimizing ECM



ECM Rule

The regions that minimize the ECM are:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

If there is differential prevalence:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

If there is differential cost:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

And if we have both differential prevalence and differential cost:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}$$

Two normal populations with equal covariance matrices

For continuous **X**, we assume multivariate normality:

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)}$$

$$f_2(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_2)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)}$$

Two-group linear discriminant analysis

Sample based ECM Rule: assign observation $\mathbf{x}$ to population 1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left(\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right)$$

where $\mathbf{S}_p$ is the pooled covariance matrix:

$$\mathbf{S}_p = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$$

Two-group linear discriminant analysis

Define:

$$\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \qquad y = \mathbf{a}'\mathbf{x}$$

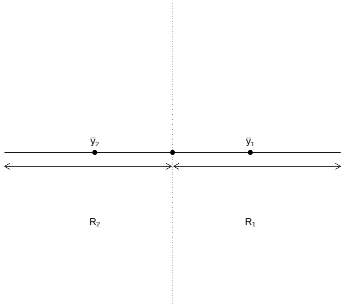
Note that:

$$y_i = \mathbf{a}'\mathbf{x}_i \qquad \overline{y}_1 = \mathbf{a}'\bar{\mathbf{x}}_1 \qquad \overline{y}_2 = \mathbf{a}'\bar{\mathbf{x}}_2$$

With equals costs and priors, the ECM rule for R_1 boils down to the univariate rule:

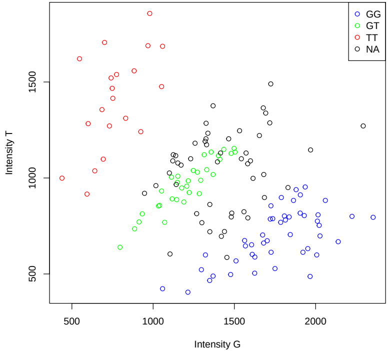
$$y_i > \frac{1}{2}(\overline{y}_1 + \overline{y}_2)$$

y is the classifier or linear discriminant function.



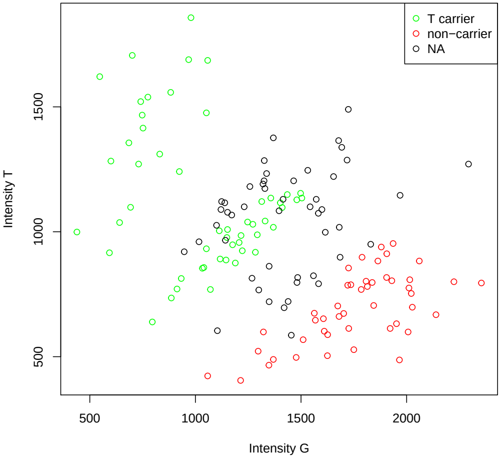
Example: SNP intensities and called genotypes

	SNP	iG	iT
1	TT	641	1037
2	GT	1207	957
3	TT	1058	1686
4	GG	1348	466
5	GT	1176	948
6	GG	1906	912
⋮	⋮	⋮	⋮
12	NA	947	920
⋮	⋮	⋮	⋮



- Calling algorithm assigns missings to "difficult" genotypes
- Could we reasonably predict if these are carriers of the T allele?

Re-plotting


T carrier
non−carrier
NA
Intensity T
Intensity G
500
1000
1500
2000

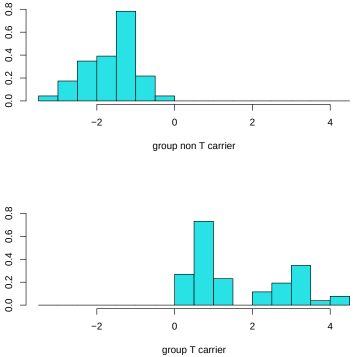
Two-group LDA output

Model: Carrier status $\sim$ iT + iG

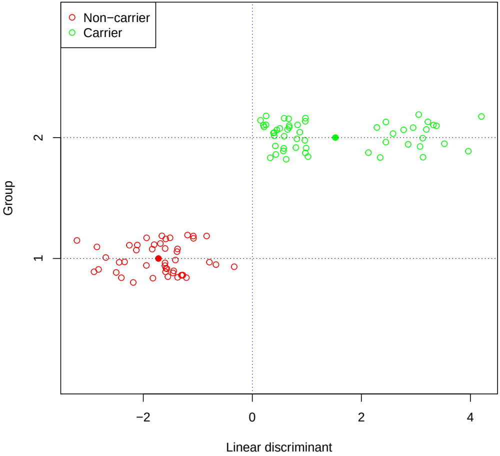
		group means	
	Prior	iT	iG
non T carrier	0.47	691.59	1758.78
T carrier	0.53	1133.56	1037.44

Linear discriminant function:

$$\text{LD1} = 0.002525858 \text{ iT } -0.002951084 \text{ iG}$$



Graphical representation



Prediction of missings

			Posterior prob.	
	Prediction	LD1	non.T.carrier	T.carrier
12	T carrier	1.25	0.01	0.99
20	non T carrier	-0.25	0.59	0.41
21	non T carrier	-0.98	0.94	0.06
27	T carrier	1.15	0.02	0.98
28	T carrier	0.22	0.24	0.76
29	non T carrier	-1.83	1.00	0.00
⋮	⋮	⋮	⋮	⋮

Two-group QDA

Under the assumption of multivariate normality with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, using the same ECM principle, a quadratic classification rule is obtained.

Sample based ECM Rule: assign observation $\mathbf{x}$ to population 1 if

$$-\frac{1}{2}\mathbf{x}'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x} + (\overline{\mathbf{x}}_1{}'\mathbf{S}_1^{-1} - \overline{\mathbf{x}}_2{}'\mathbf{S}_2^{-1})\mathbf{x} - k \geq \ln\left(\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right)$$

with

$$k = \frac{1}{2}\ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right) + \frac{1}{2}(\overline{\mathbf{x}}_1'\mathbf{S}_1^{-1}\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2'\mathbf{S}_2^{-1}\overline{\mathbf{x}}_2)$$

Sample covariance matrices

Non-carriers		
	iG	iT
iG	71677.24	24891.89
iT	24891.89	20914.78

Carriers		
	iG	iT
iG	77553.35	-23117.55
iT	-23117.55	81695.66

Predictions for missings

	LDA				QDA		
	Prediction	LD1	non.T.carrier	T.carrier	Prediction	non.T.carrier.1	T.carrier.1
12	T carrier	1.25	0.01	0.99	T carrier	0.00	1.00
20	non T carrier	-0.25	0.59	0.41	T carrier	0.00	1.00
21	non T carrier	-0.98	0.94	0.06	non T carrier	0.79	0.21
27	T carrier	1.15	0.02	0.98	T carrier	0.00	1.00
28	T carrier	0.22	0.24	0.76	T carrier	0.00	1.00
29	non T carrier	-1.83	1.00	0.00	non T carrier	0.99	0.01
32	T carrier	0.10	0.31	0.69	T carrier	0.00	1.00
35	non T carrier	-0.64	0.83	0.17	non T carrier	0.54	0.46
41	T carrier	0.87	0.04	0.96	T carrier	0.00	1.00
47	T carrier	0.83	0.04	0.96	T carrier	0.00	1.00
48	T carrier	0.99	0.02	0.98	T carrier	0.00	1.00
52	T carrier	0.04	0.36	0.64	T carrier	0.03	0.97
58	non T carrier	-0.95	0.93	0.07	non T carrier	0.84	0.16
62	non T carrier	-0.52	0.78	0.22	T carrier	0.09	0.91
65	non T carrier	-0.80	0.90	0.10	non T carrier	0.69	0.31
69	non T carrier	-0.71	0.87	0.13	non T carrier	0.74	0.26
72	non T carrier	-0.70	0.86	0.14	non T carrier	0.71	0.29
75	non T carrier	-0.67	0.85	0.15	T carrier	0.17	0.83
76	T carrier	1.24	0.01	0.99	T carrier	0.00	1.00
80	non T carrier	-0.18	0.53	0.47	T carrier	0.13	0.87
81	T carrier	0.44	0.13	0.87	T carrier	0.00	1.00
83	T carrier	-0.08	0.45	0.55	T carrier	0.00	1.00
87	T carrier	-0.01	0.40	0.60	T carrier	0.23	0.77
89	T carrier	0.35	0.17	0.83	T carrier	0.00	1.00
92	T carrier	1.04	0.02	0.98	T carrier	0.00	1.00
95	non T carrier	-1.28	0.98	0.02	non T carrier	0.93	0.07
101	T carrier	1.07	0.02	0.98	T carrier	0.00	1.00
102	T carrier	0.89	0.03	0.97	T carrier	0.00	1.00
104	T carrier	0.96	0.03	0.97	T carrier	0.00	1.00
106	T carrier	1.18	0.01	0.99	T carrier	0.00	1.00
108	non T carrier	-1.09	0.96	0.04	non T carrier	0.92	0.08
110	T carrier	1.05	0.02	0.98	T carrier	0.00	1.00
115	T carrier	0.40	0.15	0.85	T carrier	0.00	1.00
118	non T carrier	-1.19	0.97	0.03	non T carrier	0.68	0.32
121	T carrier	1.16	0.01	0.99	T carrier	0.00	1.00
122	T carrier	-0.08	0.46	0.54	T carrier	0.03	0.97
123	non T carrier	-0.59	0.82	0.18	T carrier	0.45	0.55
126	T carrier	0.34	0.18	0.82	T carrier	0.00	1.00
127	T carrier	0.79	0.05	0.95	T carrier	0.00	1.00
128	T carrier	0.85	0.04	0.96	T carrier	0.00	1.00
129	T carrier	-0.10	0.47	0.53	T carrier	0.00	1.00
131	T carrier	0.40	0.15	0.85	T carrier	0.00	1.00
134	T carrier	-0.05	0.43	0.57	T carrier	0.00	1.00
135	T carrier	-0.06	0.44	0.56	T carrier	0.00	1.00
138	T carrier	1.16	0.01	0.99	T carrier	0.00	1.00
139	T carrier	0.76	0.05	0.95	T carrier	0.00	1.00
144	non T carrier	-0.44	0.73	0.27	T carrier	0.44	0.56
145	non T carrier	-0.23	0.58	0.42	T carrier	0.00	1.00

Error rates and Confusion matrix

- It is of interest to evaluate the performance of a classification rule.
- There are several criteria to do so.
- Actual error rate (AER, density dependent)

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

- Apparent error rate (APER, not density dependent) based on the confusion matrix

		Predicted class	
		π_1	π_2
True	π_1	n_{11}	n_{12}
Class	π_2	n_{21}	n_{22}

- APER obtained as

$$\text{APER} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

- APER underestimates the AER.

Error rates and Confusion matrix

- It is of interest to evaluate the performance of a classification rule.
- There are several criteria to do so.
- Actual error rate (AER, density dependent)

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

- Apparent error rate (APER, not density dependent) based on the confusion matrix

		Predicted class	
		π_1	π_2
True	π_1	n_{11}	n_{12}
Class	π_2	n_{21}	n_{22}

- APER obtained as

$$\text{APER} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

- APER underestimates the AER.

Jackknife or hold-one-out

Procedure:

- Take the data from group π_1. Omit the ith observation, build the classifier with $n_1 - 1 + n_2$ observations.
- Classify the ith observation using the classifier.
- Repeat for all observations in π_1.
- Calculate n_{1M}^H, the number of observations that were held out and misclassified.
- Do the same for group π_2 and calculate n_{2M}^H.
- Obtain an estimate of the expected actual error rate

$$E\left(\text{AER}\right) = \frac{n_{1M}^H + n_{2M}^H}{n_1 + n_2}$$

Allele intensities revisited

LDA

	non T carrier	T carrier
non T carrier	46	0
T carrier	0	52

$$\text{APER} = \frac{0+0}{46+52} = 0$$

With cross-validation

$$\text{E(AER)} = 0$$

QDA

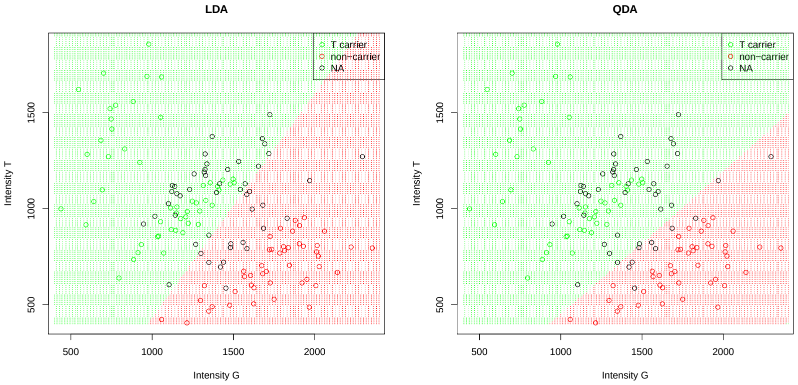
	non T carrier	T carrier
non T carrier	46	0
T carrier	0	52

$$\text{APER} = \frac{0+0}{46+52} = 0$$

With cross-validation

$$\text{E(AER)} = 0.0102$$

Visualisation



LDA with multiple groups

- The ECM rule can be extended to k groups
- Fisher's discriminant analysis

ECM rule

ECM rule with k groups (equal costs)

Assign $\mathbf{x}$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \forall \quad i \neq k$$

Fisher's linear discriminant analysis

- Searches for an optimal linear combination:

$$Z_1 = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

- Maximizes the ratio of variability between groups to variability within groups
- Objective function

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

- $\mathbf{W}$ is the matrix with within-group sums-of-squares
- For a single group i

$$\mathbf{W}_i = (\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')'(\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')$$

- $\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$
- $\mathbf{B}$ is the matrix with between-group sums-of-squares
- $\mathbf{T}$ is the matrix with total sums-of-squares

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}\mathbf{m}')'(\mathbf{X} - \mathbf{1}\mathbf{m}') \qquad \mathbf{T} = \mathbf{W} + \mathbf{B}$$

Solution

- The optimal weights are found by solving an eigenvector-eigenvalue problem

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

- The number of dimensions d in the solution is given by $\min(k-1, p)$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{D}_{\lambda}$$

- Eigenvectors scaled to satisfy $\mathbf{A}'\mathbf{S}_p\mathbf{A} = \mathbf{I}$
- Selecting the first two eigenvalues and eigenvectors allows for dimension reduction

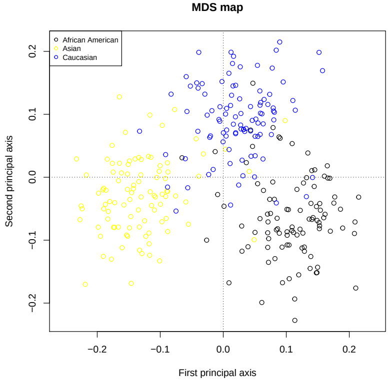
NIST autosomal STR data revisited

The data:

- 29 autosomal STRs
- Consider individuals with African-American, Asian and Caucasian ancestry
- Sample sizes balanced by subsampling

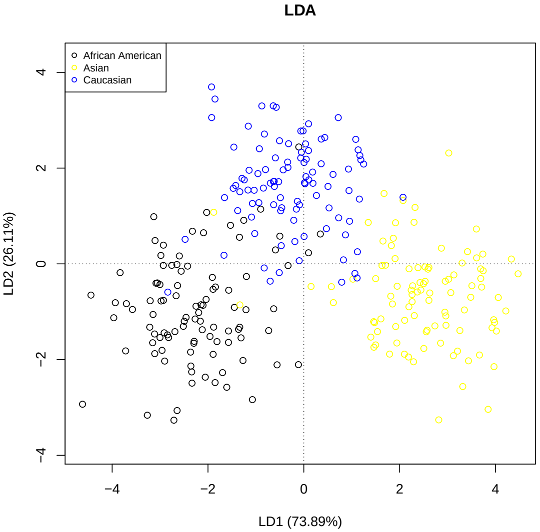
Prior to discriminant analysis:

- STRs coded as binary variables
- Quantification of the data by MDS based on Jaccard metric



Can we predict ancestry from an STR profile?

STR data in discriminant space



Numerical output

	1	2
Eigenvalue	550.38	194.45
Fraction	0.74	0.26
Cumulative	0.74	1.00

		Principal axis									
	prior	1	2	3	4	5	6	7	8	9	10
Afr. Ame.	0.333	0.108	-0.063	-0.001	-0.001	0.002	0.001	0.002	0.010	0.010	0.009
Asian	0.333	-0.132	-0.029	0.007	-0.002	0.010	0.005	-0.007	-0.002	-0.011	0.003
Caucasian	0.333	0.024	0.092	-0.006	0.003	-0.012	-0.006	0.006	-0.009	0.002	-0.012

Confusion matrix

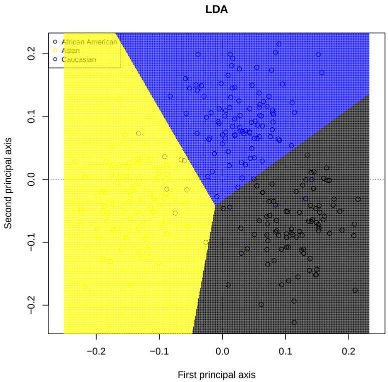
LDA			
	Afr. Ame.	Asian	Caucasian
Afr. Ame.	86	0	11
Asian	1	92	4
Caucasian	5	6	86

APER $= 0.093$

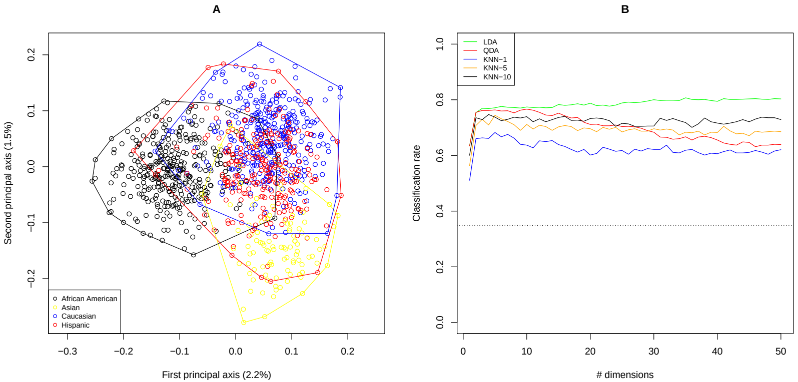
QDA			
	Afr. Ame.	Asian	Caucasian
Afr. Ame.	91	0	6
Asian	2	93	2
Caucasian	5	3	89

APER $= 0.062$

NIST STR data revisited



More complex...



Alternative statistical techniques

- An alternative technique for two-group DA is logistic regression
- An alternative technique for multi-group DA is the multinomial logit model

References

- Hand, D.J. (1981) Discrimination and Classification. Wiley, New York.
- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Chapter 11.
- Lachenbruch, P.A. (1975) Discriminant Analysis. Hafner Press, New York.